VISION PAPER

Open Science—A Question of Trust

Jonathan Clark†

The DOI Foundation, c/o EDItEUR, London N7 9DP, United Kingdom

Keywords: Persistent identifiers; Trust; Open Science; Metadata

Citation: Clark, J.: Open Science—A question of trust. Data Intelligence 3(1), 64-70 (2021). doi: 10.1162/dint_a_00078

ABSTRACT

Collaboration and the sharing of knowledge is at the heart of Open Science (OS). However, we need to know that the knowledge we find and share is really what it purports to be; and we need to know that the authors we hope to collaborate with are really the people they claim to be. In this paper, the author argues that a prerequisite for OS is trust and that persistent identifiers help to build that trust. The persistent identifier systems must themselves be trustworthy and they must be able to connect the user or their machine to the information they need now and into the future. Infrastructure is rather like plumbing: It goes unnoticed and unappreciated until it fails. This paper puts infrastructure for persistent identifiers in the spotlight as a core component of OS.

1. INTRODUCTION

Open Science (OS) is the movement to make scientific research and data accessible to all, making it easier to publish and communicate scientific knowledge and to make science more transparent and accessible during the research process [1]. For it to be successful there must be trust in the information that is made accessible to all and trust in the infrastructure that supports the exchange of information. OS is, by its very nature, a distributed system and thus a way must be found to build and manage trust across a distributed information network.

2. TRUST

Trust is the (reasonably firm) belief that you will behave in such a way towards someone even when it is not in your immediate interest to do so [2]. In other words, when we say we trust someone we are

† Corresponding author: Jonathan Clark (Email: jonathanmtclark@gmail.com; ORCID: 0000-0001-9551-9662).

implicitly saying that the chance of that person behaving in a beneficial way is high enough for us to consider engaging in some form of cooperation with them.

In an OS architecture there are multiple layers of trust. We must know if we can trust the data and if we can trust the interpretation of that data. However, before we can reach that layer of trust, we need to know that the information we have found is really what it purports to be, and that the authors are really the people they claim to be. In other words, we need to know reliably over time what's what and who's who. This is exactly what a persistent identifier does. A persistent identifier (or PID) is a long-lasting reference to a digital resource. An identifier is a label which gives a unique name to an entity: a person, place, or thing. A persistent identifier reliably points to a digital entity. Persistent identifiers are an essential part of an OS architecture [3,4].

It follows that we need a shared infrastructure for persistent identifiers, one that can be trusted and relied on by the OS community. Bilder et al. [5] identify three areas necessary for an organizations to establish a trusted shared infrastructure: running the infrastructure (governance), funding it (sustainability), and preserving community ownership of it (insurance).

Governance by a board of stakeholders helps to ensure that the organization does not confuse serving itself with serving its stakeholders [5]. Trust in the governance can be built through openness and transparency.

Financial stability is key to creating trust. As Bilder et al. note, even a well-meaning organization will not be trusted if it does not have sustainable resources.

Insurance is needed to maintain long-term trust. Should the organization fail at some point in the future, however unlikely that may seem, the shared infrastructure must be able to persist in some functional form. In practice, this is very challenging. Paskin [6] describes the DOI® system as both a technical and a social infrastructure. Whilst there may be ways to secure the technical aspects (e.g., open source, software escrow, dark archives), re-establishing the social infrastructure after a failure would be much less straightforward.

The Research Data Alliance (RDA) developed criteria for trustworthy PID Systems [7] as shown in Figure 1.

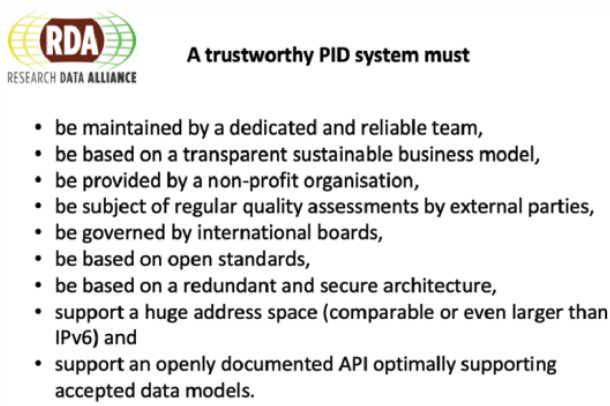

Figure 1. RDA criteria for trustworthy PID systems.

3. CONNECTIONS

Persistent identifiers that support and enable OS must be globally unique, resolvable and persistent [3]. To achieve this, the identifiers must have a syntax specification to avoid clashes, a resolution component providing the mechanism to resolve the identifier to something, metadata to bind the identifier to the thing being identified, and the technical and social infrastructure to guarantee persistence.

Put simply, in order to be useful a persistent identifier must not only answer the questions of what is what and who is who but also be able to connect the user or their machine to the information they need. As noted by Meadows et al. [4], persistent identifiers enable clear, reliable and unambiguous connections between people, places and things, and help make science more easily discoverable.

Linking scientific articles through the references (citation linking) is already well established. However a solid and sustainable persistent identifier infrastructure allows for the creation of a much richer set of links between digital objects.

The assertion of a defined link between two objects has intrinsic value and can contribute to the advancement of science. For example, researchers saw a connection between the flight of the wandering albatross and chaos theory that has profound implication for animal foraging and efficient routing in a network [9].

The Scholix Initiative (Scholarly Link Exchange) has established a framework for link exchange [10]. It was developed to facilitate the linking between journal articles and data sets. The current schema defines the following relationship types: IsSupplementTo; IsSupplementedBy; References; IsReferencedBy; IsRelatedTo. However, the schema could be extended to describe any relationship between a source and a target object.

The Entertainment Identifier Registry (which is an identifier system for audio visual material) has a concept of Alternate IDs (in essence a SameAs or AlsoKnownAs attribute). This allows objects that have multiple identifiers to be mapped to each other and disambiguated [11,12].

ORCID provides a persistent digital identifier for researchers. Using the ORCID ID to connect records across other identifier systems, ORCID is able to provide researchers with the ability to auto-update their records with their work and for trusted organizations to update their records on their behalf [13].

Crossref supports the following relationships: is-review-of, is-preprint-of, is-identical-to, has-preprint, is-comment-on, is-reply-to, has-review, has-reply, and is-supplemented-by.

There is some, albeit empirical, evidence to show that trust is transferred through links and connections, even across different contexts [14,15]. It follows that establishing a network of links between trusted information sources would strengthen an OS infrastructure. Interoperability is a worthy goal even if it is not always easy to achieve in practice.

4. METADATA

One of the first applications of a persistent identifier system for science was to enable reference linking between journal articles [16]. The importance of metadata was recognized at the start but seen as only something necessary to support reference linking. After all, the valuable item was the journal article content and the insights contained therein. Since those early days in 1999, around 80 million journal articles have been assigned persistent identifiers [17] and a vast amount of metadata has been collected.

It has become clear over the years that when there is a large enough set of metadata, it has a value all of its own. Moreover, metadata can be added over time even though the content of the journal article remains unchanged. An excellent example of this is the Crossref Funder Registry, which is an open registry of grant-giving organization names and identifiers. It allows funder information to be added to the metadata so that anyone can make connections, for example, to identify which funders invest in certain fields of research. Funding data are also used by funders to track the publications that result from their grants.

Data provenance metadata is the information on why and how the data were produced, where, when and by whom the data were collected, and who, where and when the data have been used or referred to. The role of provenance in building trust has received a lot of attention recently, most of it related in some way to blockchain [18]. However, the concept that knowing where something originated from and how it came to be, is in no way limited to blockchain. Indeed, registry-based technologies such as the Handle System [19] are well suited to provenance tracking.

5. CURATION

Gambetta [2] asks “can we trust trust?” and argues that trust is rarely based on evidence but on the lack of contrary evidence. Distrust, he says, has the capacity to be self-fulfilling. It follows that a trusted persistent identifier organization needs to make sure the elements of the infrastructure they provide work the way they were intended. Most especially they need to be able to correct errors and fix broken links.

Broken links result from broken binding between the identifier and the object being identified. This can be caused by the object being moved to another location or the disappearance of the object altogether.

Persistent identifiers must have a level of abstraction such that the new location of an object can be updated without the identifier itself having to change. When a user or their machine resolves the identifier, they are redirected seamlessly to the location. Similarly, if there are errors in the metadata, then these can be corrected without the identifier itself having to change.

The situation when the object itself disappears and cannot be found anywhere is harder to deal with. Weigl et al. [20] have proposed that persistent identifier systems should support the concept of Kernel Information. This is a minimum metadata set that enables an object to be recognized and interpreted by

machines. Resolution to the Kernel Information must still be possible when an object is no longer available. Another option is to create a "tombstone" with helpful information drawn from the metadata of the object that has disappeared, including its last known location [6].

6. CHOOSING A PID SYSTEM

There are many PID systems to choose from: Archival Resource Keys (ARKs), Digital Object Identifiers (DOIs), Handles, Uniform Resource Names (URNs) to name but a few. The Dutch Digital Heritage Network developed a self-assessment tool called the PID Guide to form and guide organizations in their choices [21]. The European Open Science Cloud (EOSC) initiative has produced a PID Policy which defines a set of expectations about persistent identifiers in support of FAIR research [21]. In the end, the choice of which PID system to adopt comes down which system is trusted the most. For this reason, the key first question to ask when choosing a system is: to which community do we belong or feel the most affinity for?

Having said that, there is a growing diversity of use cases for PIDs and it is inevitable that multiple PID systems will need to co-exist. What is important for users is that there is unambiguous linking between persistent identifiers that identifies the same type of object, for instance: journal articles citing other journal articles. The FREYA project has developed a PID Graph as a way to connect existing persistent identifiers to each other in standardized ways [22,23].

7. FINAL WORDS

Persistent identifiers enable researchers, their organizations and their research to be uniquely identified and connected. Knowing what is what and who is who is fundamental for the research information infrastructure. It is worth noting that infrastructure is rather like plumbing. It goes unnoticed and unappreciated until it fails. The business of building and maintaining persistent identifier systems is especially challenging for this reason. Yet is could not be more important for the success of OS.

REFERENCES

[1] UNESCO global open access portal. Available at: http://www.unesco.org/new/en/communication-and-information/portals-and-platforms/goap/open-science-movement/. Accessed 7 January 2021
[2] Gambetta, B.: Trust: Making and breaking cooperative relations. Blackwell, Oxford (1988)
[3] Valle, M., et al.: A Persistent Identifier (PID) policy for the European Open Science Cloud, Report from the European Open Science Cloud FAIR and Architecture Working Groups. Available at: https://doi.org/10.2777/926037. Accessed 7 January 2021
[4] Meadows, A., Haak, L.L., Brown, J.: Persistent identifiers: The building blocks of the research information infrastructure. Insights 32(1), 9 (2019)
[5] Bilder, G., Lin, J., Neylon, C.: Principles for open scholarly infrastructure-v1. Figshare (2015). Available at: https://doi.org/10.6084/m9.figshare.1314859. Accessed 7 January 2021

[6] Paskin, N.: 2009, Digital Object Identifier (DOI®) System. In: Bates, M.J., Maak, M.N. (eds.) Encyclopedia of Library and Information Sciences (3rd edition), pp. 1–10. CRC Press, Boca Raton (2009)
[7] Wittenburg, P., et al.: Persistent identifiers: Consolidated assertions. Status of November, 2017. Available at: http://doi.org/10.5281/zenodo.1116189. Accessed 7 January 2021
[8] Caplan, P., Arms, W.: Reference linking for journal articles. D-Lib Magazine 5(7/8). Available at: https://doi.org/10.1045/july99-caplan. Accessed 7 January 2021
[9] Viswanathan, G., et al.: Lévy flight search patterns of wandering albatrosses. Nature 381, 413–415 (1996)
[10] Cousijn, H., et al.: Bringing citations and usage metrics together to make data count. Data Science Journal 18(1), 9 (2019)
[11] Kroon, R.: Bringing order to digital identifiers. Media and Entertainment Journal Winter, 148-150 (2014)
[12] Peña, J., Dulchinos, D. EIDR ID round trip validates vision of unique identifiers. Media and Entertainment Journal Spring, 54 (2017)
[13] Akers, K.G, et al.: ORCID author identifiers: A primer for librarians. Medical Reference Services Quarterly, 35(2), 135–144 (2016)
[14] Stewart, K.J.: Trust transfer on the World Wide Web. Organization Science 14(1), 5–17 (2003)
[15] Buntain, C., Golbeck, J.: Trust transfer between contexts. Journal of Trust Management 2, Article number 6 (2015)
[16] Atkins, H., et al. 2000. Reference linking with DOIs: A case study. D-Lib Magazine 6(2) (2000). Available at: https://doi.org/10.1045/february2000-risher. Accessed 7 January 2021
[17] Crossref Annual Report & Fact File 2018–19. Available at: https://doi.org/10.13003/y8ygwm5. Accessed 7 January 2021
[18] Montecchi, M., Plangger, K., Etter, M.: It's real, trust me! Establishing supply chain provenance using blockchain. Business Horizons 62(3), 283–293 (2019)
[19] Sun, S.L., Boesch, L.B.: RFC3650 Handle system overview. Available at: https://doi.org/10.17487/RFC3650. Accessed 7 January 2021
[20] Weigel, T., et al.: RDA recommendation on PID kernel information. Research Data Alliance. Available at: https://doi.org/10.15497/RDA00031. Accessed 7 January 2021
[21] Persistent Identifier Guide. Available at: https://www.pidwijzer.nl/en. Accessed 7 January 2021
[22] A persistent identifier (PID) policy for the European Open Science Cloud, EOSC Executive Board WG FAIR and Architecture October 2020, European Commission Directorate-General for Research and Innovation. Available at: https://doi.org/10.2777/926037. Accessed 7 January 2021
[23] Fenner, M., Aryani, A.: Introducing the PID graph. Available at: https://doi.org/10.5438/jwvf-8a66. Accessed 7 January 2021

AUTHOR BIOGRAPHY

Jonathan Clark is the Managing Agent for the DOI Foundation (which is a not-for-profit membership organization that governs the DOI (Digital Object Identifier) and is the registration authority for the ISO standard (ISO 26324). The DOI system provides a technical and social infrastructure for the registration and use of persistent identifiers called DOIs. Jonathan also works as an independent advisor on strategy and innovation. He is a Guest Lecturer and External Examiner for the Masters in Imagineering program at the Breda University of Applied Sciences. Prior to this he worked at Elsevier for 20 years in various positions in publishing, marketing and technology. He holds a B.Sc. and PhD in Chemical Engineering from the University of Newcastle-upon-Tyne. Jonathan was Chair and Director of the DOI Foundation from 2005-2010. He lives in Croatia.

ORCID: 0000-0001-9551-9662